Proposed Task Description for

Knowledge-Base Population at TAC 2014

English Sentiment Slot Filling

May 1, 2014

1. Introduction

The goal of Sentiment Slot Filling is to use corpora to collect information regarding sentiment expressed towards or by an entity. Sentiment is defined here as *a positive or negative emotion, evaluation, or judgement.* Entities may be a person (PER), organization (ORG), or a geopolitical entity (GPE). This task explores the sentiment triple:

<sentiment holder, sentiment, sentiment target>

Which we formalize for purposes of evaluation as:
{query entity, sentiment slot} → filler entity

Queries will include a query entity and a sentiment slot that indicates both query polarity and directionality. Thus, depending on the sentiment slot, the query entity will either be a sentiment holder or a sentiment target. Systems must return unique values for the remaining member of the triple: either sentiment targets or sentiment holders, depending on the sentiment slot.

For example, if the query specifies an entity with positive polarity towards X, systems must return distinct entities towards which the query entity holds a positive sentiment (the sentiment targets). If the query specifies an entity with negative polarity from X, systems must return distinct entities that hold negative sentiment towards the query entity (the sentiment holders). Possible answers therefore fill one of the following slots:

> **pos-towards:** query entity holds positive sentiment towards filler entity ("likes", "is hopeful about", etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment targets*.
> **pos-from:** query entity is target of positive sentiment from filler entity ("is liked by", "was hoped for by", etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment holders.*
> **neg-towards:** query entity holds negative sentiment towards filler entity ("dislikes", "is skeptical about", etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment targets*.
> **neg-from:** query entity is target of negative sentiment from filler entity ("is disliked by", etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment holders.*

Sentiment may be directed toward an entity based on direct evaluation of an entity (e.g. "Kentucky doesn't like Mitch McConnell") or may be directed to an entity based on actions that the entity took (e.g. "Kentucky doesn't like Mitch McConnell's stance on gun control"). Given a query with {Mitch McConnell, neg-from}, the filler would be the holder of the

sentiment, Kentucky.

For this task, sentiment slots are understood to be invertible: If a filler entity Y is correct for a query {X, pos-from}, then the filler entity X is correct for a query {Y, pos-towards}.

Automatically annotated named entity and within-document coreference chains from BBN's SERIF system is provided for all of the documents in the TAC 2014 KBP English Source Corpus (LDC2014E13). Note that this information is just a supportive tool. LDC is not restricted to the entity and coreference chains in sentiment slot filling query development and assessment, and systems are not restricted to the entity and coreference chains in sentiment slot filling. Sentiment expressed towards or by entities that are not PER, ORG, or GPE are not correct slot fillers for this task, even when incorrectly labelled as such in the provided annotations.

Sentiment holders and targets may also be taken from metadata, as in the case of post authors for discussion forum data. (For an example, see "Query Development".)

2. Input

Each query in the Sentiment Slot Filling task consists of a query ID, the name of the entity, a document (from the TAC 2014 KBP English Source Corpus) in which the name appears, the start and end offsets of the name as it appears in the document (to disambiguate the query entity in case there are multiple entities with the same name), the query entity's type (PER, ORG, or GPE), and the sentiment slot to be filled (which specifies whether the query entity is a sentiment holder or a sentiment target, and the polarity of the sentiment held by or about them). An example query is:

```
<query id="SSF_ENG_002">
  <name>PhillyInquirer</name>
  <docid>eng-NG-31-141808-9966244</docid>
  <beg>757</beg>
  <end>770</end>
  <enttype>ORG</enttype>
  <slot>pos-towards</slot>
</query>
```

3. Output

The output should be a set of distinct slot filler entities for the query, *that is found in the same document as the query*. For the example SSF_ENG_002 query above, the system should output each distinct person, organization, or GPE that the entity PhillyInquirer has expressed a positive sentiment towards in the document eng-NG-31-141808-9966244.[1]

3.1 Format

System output files should be in UTF-8 and contain a *ranked set of responses* for each query id, with highest scores indicating highest confidence. LDC reserves the right to limit assessment to top responses. A response consists of a single line, with a separate line for each distinct slot filler. Lines should have the following tab-separated columns:

[1] Restricting slot fillers to be attested in the same document as the query is a change from the TAC 2013 Sentiment slot

Column 1	Query id
Column 2	a sentiment slot name, indicating polarity and directionality (same as in the query)
Column 3	A unique run id for the submission
Column 4	NIL, if the system believes that no information is learnable for this slot, in which case Columns 5-7 are empty; or provenance for the relation between the query entity and slot filler, consisting of up to ***4 triples*** in the format: docid:startoffset-endoffset separated by comma. The docid must be for the same document as was provided in the query. Each of the individual spans may be at most ***150 UTF-8 characters***.
Column 5	A slot filler string (name of an entity)
Column 6	Provenance for the slot filler string. This is either a single span (docid:startoffset-endoffset) from the document where the slot filler string was extracted, or (in the case when the slot filler string in Column 5 has been normalized, e.g., "Bill Clinton" extracted and normalized from "Bill and Hillary Clinton") a set of up to two docid:startoffset-endoffset spans for the base strings that were used to generate the normalized slot filler string. As in Column 4, multiple spans must be separated by commas. The document used for the slot filler string provenance must be the same as the document in Column 4. LDC will judge Correct vs. Inexact with respect to the document provided in the slot filler string provenance.
Column 7	Confidence score (used to rank the responses)

Sample output:

Column 1: SSF_ENG_002
Column 2: pos-towards
Column 3: TeamX5
Column 4: APW_ENG_20101231.0403:521-800
Column 5: Washington Post
Column 6: APW_ENG_20101231.0403:549-563
Column 7: 0.6

The output file should contain a separate line for each unique slot filler entity returned for a query. When no information is believed to be learnable for a slot, Column 4 should be NIL and Columns 5-7 should be left empty. Column 5 contains a string representing the slot filler entity -- a person (PER), organization (ORG) or geopolitical entity (GPE); the string should be extracted from the document in Column 4, except that any embedded tabs or newline characters should be converted to a space character.

Provenance: The provenance stored in Column 4 must contain text that justifies the extracted relation. That is, it must include *some* mention of the query and slot filler entities and some text supporting the sentiment relation between them. The spans must contain a mention of the slot filler that LDC can corefer with the named mention given in Column 5, and a mention of the query entity that LDC can corefer with the mention given in the query, but need not contain *named* mentions of either the query entity or slot filler.

Offsets: Each document is represented as a UTF-8 character array and begins with the "<DOC>" tag, where the "<" character has index 0 for the document. Note that the beginning <DOC> tag varies slightly across the different document genres included in the source corpus: it can be spelled both with upper case and lower case letters, and it may include additional attributes such as "id" (e.g., <doc id="doc_id_string"> is valid document start tag). Thus, offsets in Columns 4 and 6 are counted *before* XML tags are removed. In general, start offsets must be the index of the first character in the corresponding string, and end offsets must be the index of the last character of the string (therefore, the length of the corresponding mention string is endoffset – startoffset + 1).

Confidence Scores: To promote research into probabilistic knowledge bases and confidence estimation, each non-NIL response must have an associated confidence score. Confidence scores will not be used for any official TAC 2014 measure. However, the scoring system may produce additional measures based on confidence scores. For these measures, confidence scores will be used to induce a total order over the responses being evaluated; when two scores are equal, the response appearing earlier in the submission file will be considered to have a higher confidence score for the purposes of ranking. A confidence score must be a positive real number between 0.0 (representing the lowest confidence) and 1.0 (inclusive, representing the highest confidence), and must include a decimal point (no commas, please) to clearly distinguish it from a document offset.
NIST reserves the right to assess and score only the top-ranked N non-NIL responses in each submission file, where N is determined by assessing resources and the total number of responses returned by all participants.

3.2 Filler Entities

As in all the KBP slot filling tasks, the slot filler entity string (Column 5) should be the most informative named mention of the entity in the document. For example, if "William Clinton" is the only named mention of the slot filler entity in the document, then it is acceptable to return that string in Column 5; however, if "William Jefferson Clinton" is also in the document, then this more informative string should be the string in Column 5.

Sentiment slot fillers are list-valued. Multiple fillers returned for the same query should refer to distinct individuals. It is not sufficient that slot filler entity strings be distinct; they must refer to distinct individuals. For example, if the query includes {Hillary Clinton, pos-towards} (the sentiment holder is Hillary Clinton with positive sentiment towards the filler), and the system finds both "William Clinton" and "Bill Clinton" as potential fillers, it should return just one of these. Similarly, entities **should not be repeated** as slot fillers for a query: Although it is possible that Hillary Clinton may feel "pos-towards" William Jefferson Clinton on many separate occasions, systems should only return *one* of these instances as a response.[2]

To aid in this task, we will provide automatic standoff coreference chains and named entity tags for source documents, provided by BBN's SERIF system.

For further information on what is expected in output, please see the Assessment Guidelines available here: http://www.nist.gov/tac/2014/KBP/Sentiment/guidelines.html.

[2] However, it is permitted that "William Jefferson Clinton" be the correct answer for two separate queries involving the same query entity, but different sentiment slots, e.g., for a {Hillary Clinton, pos-towards} query and a {Hillary Clinton, neg-towards} query (with appropriate justification, as usual).

4. Query Development

The source data for sentiment slot filling query development and sentiment slot filling runs will be English newswire, web text, and discussion forum data. For the discussion forum corpora, post authors may be given as query entities, or returned as filler entities. Note that in the discussion forum data post authors are indicated in the post CDATA:

> <post author="publicprotector" datetime="2009-08-02T17:05:00" id="p1">
> For all the rantinmg about the success of America's actions in Iraq the reality is quite different. Millions dead, a broken country, millions of mines and ordance all over the place, mass pollution, depleted uranuim everywhere, a pupet Government in place and the country robbed by the so called victors.
> </post>

Although automatic SERIF coreference and NER chains will be provided, LDC will not be constrained to use this output during query development. In particular, because SERIF ignores text in metadata (including the post author), participants must develop their own strategies for handling query entities and slot filler entities that are mentioned in metadata.

5. Assessment

5.1 Assessment Approach

We will pool the responses from all the systems and have human assessors judge the responses. To increase the chance of including answers that may be particularly difficult for a computer to find, LDC will prepare a manual key, which will be included in the pooled responses.
The slot filler (Column 5) in each non-NIL response is assessed as Correct, ineXact, or Wrong:

1. A response that contains more than four provenance triples (Column 4) will be assessed as Wrong.
2. Otherwise, if the text spans defined by the offsets in Column 4 (+/- a few sentences on either side of each span) do not contain sufficient information to justify that the slot filler is correct, then the slot filler will also be assessed as Wrong.
3. Otherwise, if the text spans justify the slot filler but the slot filler in Column 5 either includes only part of the correct answer or includes the correct answer plus extraneous material, the slot filler will be assessed as ineXact. No credit is given for ineXact slot fillers, but the assessor will provide a diagnostic assessment of the correctness of the justification offsets for the response. LDC will judge Correct vs. Inexact with respect to the document provided in Column 6.
4. Otherwise, if the text spans justify the slot filler and the slot filler string in Column 5 is exact, the slot filler will be judged as Correct. The assessor will also provide a diagnostic assessment of the correctness of the justification offsets for the response.

Two or more system responses for the same query entity and slot may have equivalent slot fillers (i.e., refer to the same entity); in this case, the system is given credit for only one response, and is penalized for all additional equivalent slot fillers. This is implemented by assigning each correct response to an *equivalence class,* and giving credit for only one member of each class.

Although automatic coreference and NER will be provided by KBP organizers, system outputs

will be scored based on manual judgments produced during assessment about which entity mentions are coreferent (i.e., assigned to the same equivalence class).

5.2 Scoring

Given these judgments, we can count:

Correct = total number of correct equivalence classes in system responses
System = total number of non-NIL system responses
Reference = number of equivalence classes for all slots
Recall = Correct / Reference
Precision = Correct / System
F = 2*Precision*Recall / (Precision + Recall)

The F score is the primary metric for system evaluation.

6. Data

6.1 Source Document Collection

The source document collection for the KBP 2014 Sentiment Slot Filling task documents are distributed by LDC as a single corpus: "TAC 2014 KBP English Source Corpus" with Catalog ID LDC2014E13[3]. The source documents in this release comprise the following:

1. Approximately 1 millions documents from English Gigaword Fifth Edition
2. Approximately 1 million Web Documents
3. Approximately 100,000 discussion forum documents (threads)

LDC2014E13 also includes standoff coreference chains and named entity tags for each source document, provided by BBN's SERIF system.

6.2 Training and Evaluation Corpus

Table 1 summarizes the KBP 2014 training and evaluation data that we aim to provide for participants.

Corpus	Size (relations)
Training	TAC 2013 Sentiment Slot Filling training and evaluation data (LDC2013E78, LDC2013E89, LDC2013E100)
Evaluation	Approximately 100 queries for each of the 4 slots.

Table 1. Sentiment Slot Filling Data

7. External Resource Restrictions and Sharing

7.1 External Resource Restrictions

As in previous KBP evaluations, participants will be asked to make at least one run subject to certain resource constraints, primarily that the run be made as a 'closed' system: a system that does not access the Web during the evaluation period. Participants may also submit

[3] The source documents in LDC2014E13 are the same as the English documents in "TAC 2013 KBP Source Corpus" (LDC2013E45).

additional runs which access the Web. This will provide a better understanding of the impact of external resources. Further rules for both of the primary runs and additional runs are listed in Table 2.

Specific Rules	Specific Examples
Allowed	Using a Wikipedia derived resource to (manually or automatically) create training data.
	Compiling lists of name variation based on hyperlinks and redirects before evaluation.
	Using a Wikipedia-derived resource before evaluation to create a KB of world knowledge that can be used to check the correctness of facts. Note that manual annotations of this data are allowed for what is considered world-knowledge (e.g., gazetteers, lists of entities) but only automatically-generated annotations are accepted for KBs of sentiment relations that can be directly mapped to slots used in this evaluation.
	Preprocess/annotate a large text corpus before the evaluation to check the correctness of facts or aliases. As above, only automatically-generated annotations are accepted for KBs of sentiment relations that can be directly mapped to slots used in this evaluation.
Not Allowed	Using structured knowledge bases (e.g., Wikipedia infoboxes, DBPedia, and/or Freebase) to directly fill slots or directly validate candidate slot fillers for the evaluation query.
	Editing Wikipedia pages for target entities, either during, or after the evaluation.

Table 2. Rules for Using External Resources

7.2 Resource Sharing

In order to support groups that intend to focus on part of the task, participants are encouraged to share the external resources that they prepared before the evaluation. The possible resources may include intermediate results, entity annotations, parsed/SRL-/IE-annotated Wikipedia corpus, topic model features for entity linking, patterns for slot filling, etc. The sharing process can be informal (among participants) or more formal (through a central repository built by the coordinators). Please email the coordinators in order to access the central site.

8. Submissions and Schedule

8.1 Submissions

Sentiment slot filling participants will have two weeks after the evaluation queries are released to return their results. Up to five alternative system runs may be submitted by each team. Submitted runs should be ranked according to their expected score (based on development data, for example). Systems should not be modified once queries are downloaded. Details about submission procedures will be communicated to the track mailing list. The tools to validate formats will be made available at: http://www.nist.gov/tac/2014/KBP/Sentiment/.

8.2 Schedule

Please visit the sentiment slot filling website for an approximate schedule for the Sentiment Slot Filling tasks at KBP 2014: http://www.nist.gov/tac/2014/KBP/Sentiment/.

9. Mailing List and Website

The KBP 2014 website is http://www.nist.gov/tac/2014/KBP/. The website dedicated to the Sentiment slot filling task is http://www.nist.gov/tac/2014/KBP/Sentiment/. Please post any questions and comments to the list tac-kbp@nist.gov. You must be subscribed to the list in order to post to the list. Information about subscribing to the list is available at: http://www.nist.gov/tac/2014/KBP/registration.html.